Supplementary Information for "A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis" – Down et al., (2008).

Supplementary Methods:

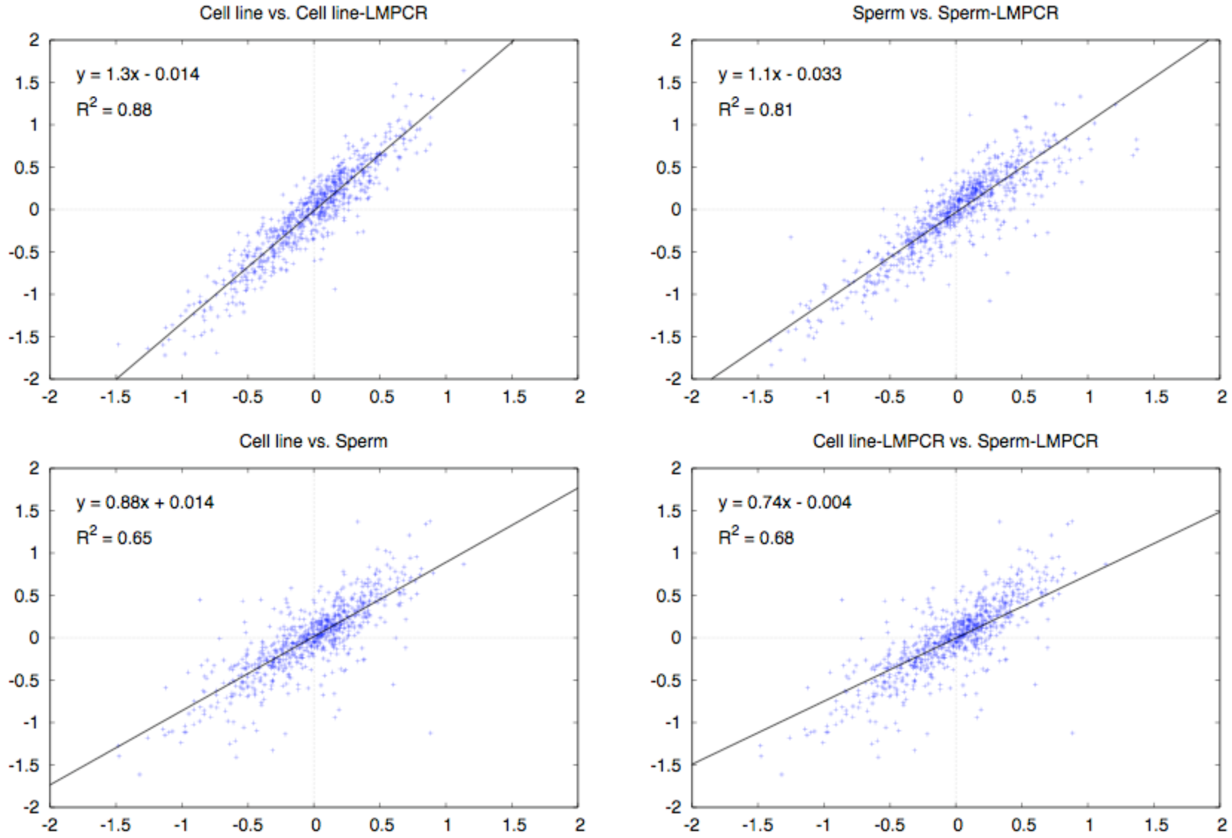
Methylated DNA Immunoprecipitation. MeDIP was based on the original protocol[1], but we included a ligation mediated PCR (LM-PCR) step[2]. Array hybridizations performed before and after LM-PCR showed that LM-PCR did not introduce significant amplification bias (**Supplementary Figure 1**). 2.5 µg of genomic DNA was sheared to a size range of 400 - 700 bp. The resulting fragments were blunt-ended by incubation for 20 minutes at 12ºC in a 120 µl reaction containing the DNA sample, 1 X Buffer 2 (NEB, U.K.), 10 X BSA (NEB, U.K.), 100 µM dNTP mix and T4 DNA polymerase (NEB, U.K.). The reaction was purified using a Zymo-5 kit (Genetix, UK) according to the manufacturer's instructions but the final elution was done in 30µl of TE buffer pH 8.5. Adapter ligation was performed by overnight incubation at 16ºC in a final volume of 100 µl containing, DNA sample, 40µl adaptors, T4 DNA ligase 10 X buffer, 5 µl T4 DNA ligase (NEB, U.K.). The reactions were purified using a Zymo-5 kit as described above. To fill in the overhangs, the DNA was incubated at 72ºC for 10 minutes in a reaction containing 100µM dNTPs, 1 X AmpliTaq Gold PCR buffer (Applied Biosystems, UK), 1.5 mM $MgCl_2$, 5U AmpliTaq Polymerase. The DNA was purified using a Zymo-5 kit as described above. 50 ng of the ligated sample was set aside as the input fraction. 1.2 µg of the ligated DNA sample was subjected to MeDIP as described previously[1], after scaling down accordingly. The immunoprecipitated (IP) sample was purified using Zymo-5 kit (using 700 µl binding buffer) according to the manufacturer's instructions. Ten nanograms of each IP and input fraction for each sample were subjected to LM-PCR using the Advantage-GC genomic PCR kit (Clontech, UK). PCR cycling conditions are available upon request. After the LM-PCR, the duplicate reactions were combined, purified using a Qiagen PCR-clean up kit (Qiagen, UK) and eluted with 50µl of water. The MeDIP and input fractions were sent to Nimblegen, Iceland for hybridization.

Comparison of pre- and post-LM-PCR on a custom MHC-tile path array (Supplementary Figs 1 and 2). Six individual Medips were performed on the GM069960 cell line (a gift from Dr Ian Dunham, Sanger Institute, UK) and sperm DNA as described above. The MHC tile-path array was constructed by the Wellcome Trust Sanger Institute Microarray Facility and will be described elsewhere (Tomazou et al., in preparation). In total 1791 ~2kb PCR clones were used to cover ~4 Mb of the human MHC. Fluorescent labeling was performed using a Bioprime labeling kit (Invitrogen) in a 130.5 µl reaction volume containing 100 ng DNA, 1.5 µl dNTP mix (2 mM dATP, 2 mM dTTP, 2 mM dGTP, and 0.5 mM dCTP), and 1.5 µl Cy5/Cy3 dCTP (1mM) (Perkin Elmer). The reactions were purified using Micro-spin G50 columns (Pharmacia-Amersham) according to the manufacturer's instructions. Reference and test samples were combined and precipitated with 3M sodium acetate (pH 5.2) in 2.5 volumes of ethanol with 90 µg human C_0t1 DNA (Invitrogen). The DNA pellet was resuspended in hybridization buffer containing 50% deionized formamide, 10% dextran sulphate, 10 mM Tris-HCl (pH 7.4), 2 × SSC, 0.1% Tween-20, and 300 µg yeast tRNA (Invitrogen). Hybridization was performed for 24 hours at 37°C on a MAUI hybridization platform. The arrays were washed serially in 2 × SSC, 0.03% SDS for 5 minutes at room temperature, and then for 5 minutes at 60°C, four times in 2 × SSC for 20 minutes at room temperature, then in PBS, 0.05% Tween20 for 10 minutes at room temperature, and finally in HPLC water for 10 minutes at room temperature. The arrays were dried and scanned using a ScanArray Express HT scanner (PerkinElmer). LOESS-normalized $\log_2$ ratios were obtained from ScanArray Express (Perkin Elmer, USA). Only clones for which the standard deviation among the 4 replicate spots was less than 0.33 were used for analysis, resulting in 750 clones.

Supplementary Table 1. Tissue samples used in the study. The "Baseline" and "Response" parameters refer, respectively, to the intercept and slope of a linear model fitted to the low-CpG portion of each array's data (refer to description of Batman in the main text). The "Response" parameter can be interpreted as the number of methylated cytosines in a region required to increase the observed array signal by one unit. Since the noise level of the arrays appears to be fairly uniform, this can be interpreted as a measure of the signal/noise ratio of the complete MeDIP-chip experiment.

Sample ID	array_id	cy3	cy5	Baseline	Response
SP2	76453	input	IP	-0.58	13.37
SP2	78923	IP	Input	-0.39	23.20
SP3	83890	IP	Input	-0.21	41.53
SP4	98489	Input	IP	-0.34	28.53

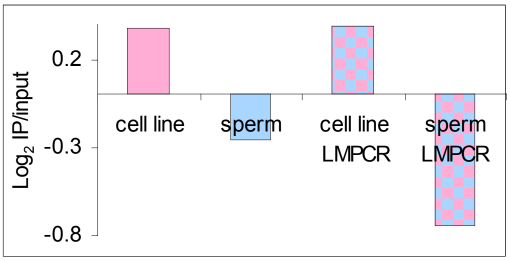
All samples were from 20 – 49 year old healthy normal males of European ancestory.



Supplementary Figure 1. Comparison of MHC-tile path array profiles pre- and post-LM-PCR. Six individual MeDIPs were performed on the GM069960 cell line (a gift from Dr Ian Dunham, Sanger Institute, UK) and sperm DNA as described above. The MHC tile-path array was constructed by the Wellcome Trust Sanger Institute Microarray Facility and will be described elsewhere (Tomazou et al., in preparation). In total 1791 ~2kb PCR clones were used to cover ~4 Mb of the human MHC. See 'Comparison of pre- and post-LM-PCR on a custom MHC-tile path array' above for description of methods. LOESS-normalized log2 ratios were obtained from ScanArray Express (Perkin Elmer, USA). Only clones for which the standard deviation among the 4 replicate spots was less than 0.33 were used for analysis, resulting in 750 clones. R^2 is a Pearson's correlation.

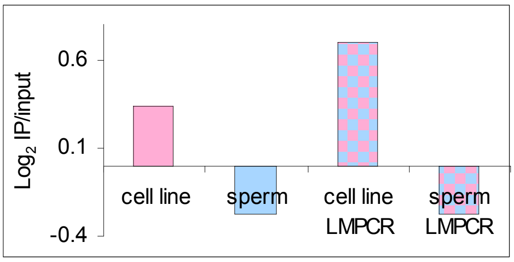
(A)

MHC tile-path array clone:	S6C_2_100k04
chromosome co-ordinates:	6: 31347352-3134994
cf.	
HEP Amplicon id:	438
chromosome co-ordinates:	6:31346417-31346805
sperm methylation:	Unmethylated (0%)
All other tissues:	Methylated
p-value	1.70E-11



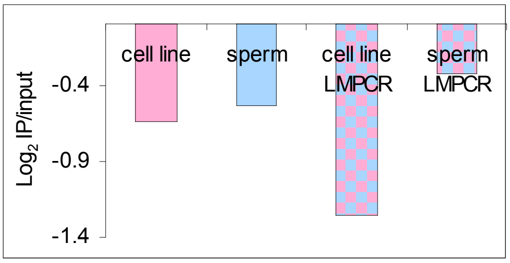
(B)

MHC tile-path array clone:	S6C_2_176p07
chromosome co-ordinates:	6: 31799483 - 31801330
cf.	
HEP Amplicon id:	414
chromosome co-ordinates:	6:31799704-31800160
sperm methylation:	Unmethylated (7%)
All other tissues:	Methylated
p-value	1.47E-04



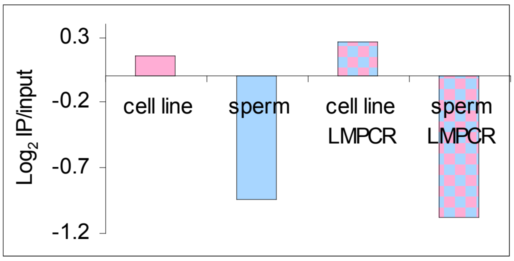
C)

MHC tile-path array clone:	S6A_2_99j22
chromosome co-ordinates:	6:33015774-33018132
cf.	
HEP Amplicon id:	443
chromosome co-ordinates:	6:33016412-33016815
sperm methylation:	Methylated (86%)
All other tissues:	Unmethylated
p-value	7.01E-08



(D)

MHC tile-path array clone:	S6A_2_182a16
chromosome co-ordinates:	6:33389687-3332295
cf.	
HEP Amplicon id:	536
chromosome co-ordinates:	6: 33391090-33391470
sperm methylation:	Unmethyalted (6%)
All other tissues:	Methylated
p-value	6.21E-18



Supplementary Figure 2. Comparison of tissue-specific Differentially Methylated Regions (tDMRs) identified from the Human Epigenome Project (HEP)[1] with the MHC tile-path array. Of the 750 clones used for further anlaysis of the MHC tile-path arrays, 4 genomic regions had been previously charcterized as tDMRs in the HEP. To test whether the MHC array has the discriminatory power to identify these tDMRs, we analysed log$_2$ ratios for these 4 regions, along with the methylation values from the HEP. The plots show these regions to be differentialy methylated based on the MHC array data, both before and after LM-PCR.

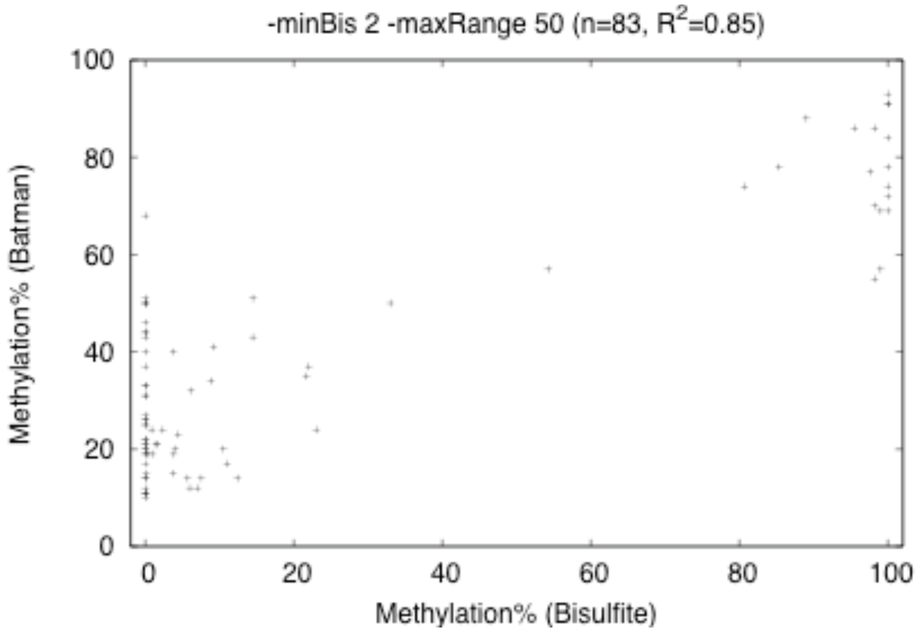
Supplementary Table 2. Batman analysis of methylated CpG islands. Batman-called MeDIP-chip methylation values of ROIs that overlap HEP amplicons classified as CpG islands in the Ensembl Genome Browser and with a mean methylation value of >80% in sperm.

HEP amplicon id[a]	chr[b]	start	end	Batman-called methylation[c]
5039	22	36,101,138	36,101,532	100
5050	22	45,308,659	45,309,008	82
5067	22	16,423,695	16,423,836	100
5224	22	20,130,812	20,131,095	96
6284	22	48,665,695	48,665,980	90
6313	22	18,510,703	18,511,128	86
6379	22	48,978,445	48,978,823	99
6385	22	48,997,725	48,998,045	89
6447	22	43,511,263	43,511,658	98
6540	22	17,498,951	17,499,329	98
6629	22	48,412,055	48,412,418	90
6662	22	25,266,974	25,267,390	90
6770	22	35,933,439	35,933,765	93
6781	22	21,768,029	21,768,430	92
13443	22	18,130,644	18,130,895	91

[a] The HEP bisulfite-PCR amplicon ids are from www.epigenome.org

[b] NCBI36 co-ordinates

[c] These are the Batman called methylation values of ROIs that overlap the corresponding HEP amplicon

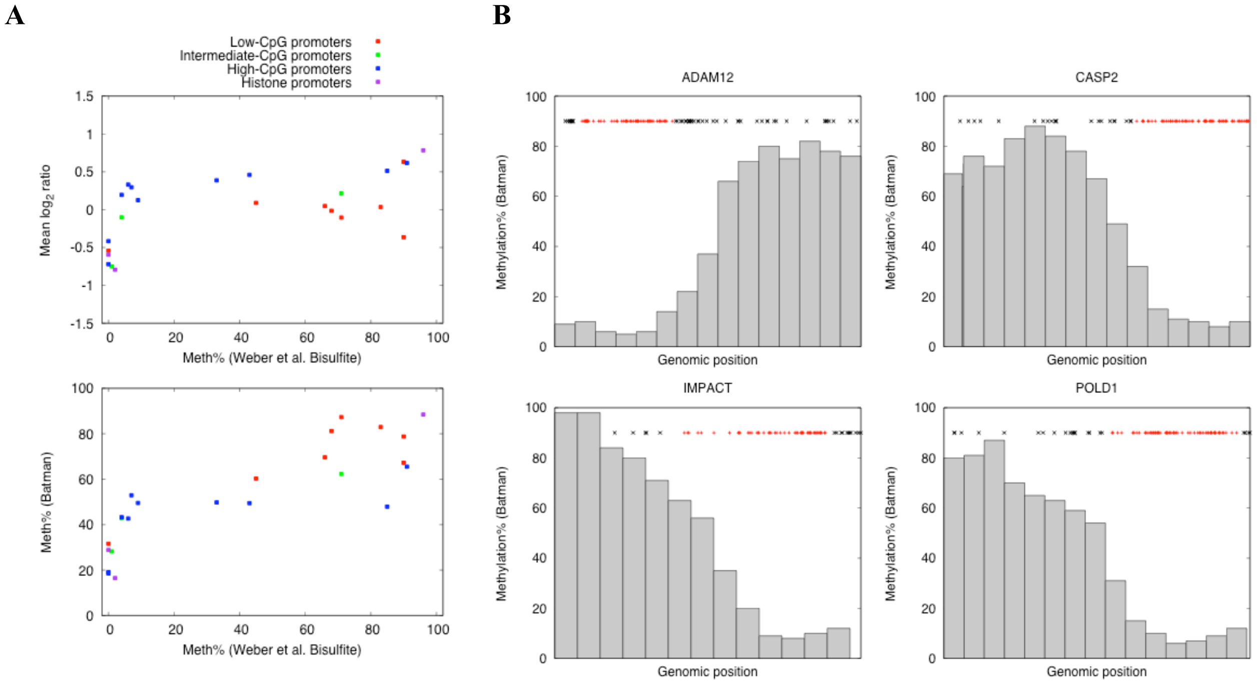


Supplementary Figure 3. Bisulfite-PCR validation of the MeDIP-chip Batman calls. Twenty-nine regions were randomly chosen for bisulfite-PCR validation, spanning a range of CpG densitites, genomic locations (see Supplementary Table 2). The validation was performed for each tissue sample used in our study. The bilsufite-PCR and data processing was performed as described previously[5,6], then averaged across 100 bp tiles (n = 83). DNA methylation data for the biological replicates were averaged.

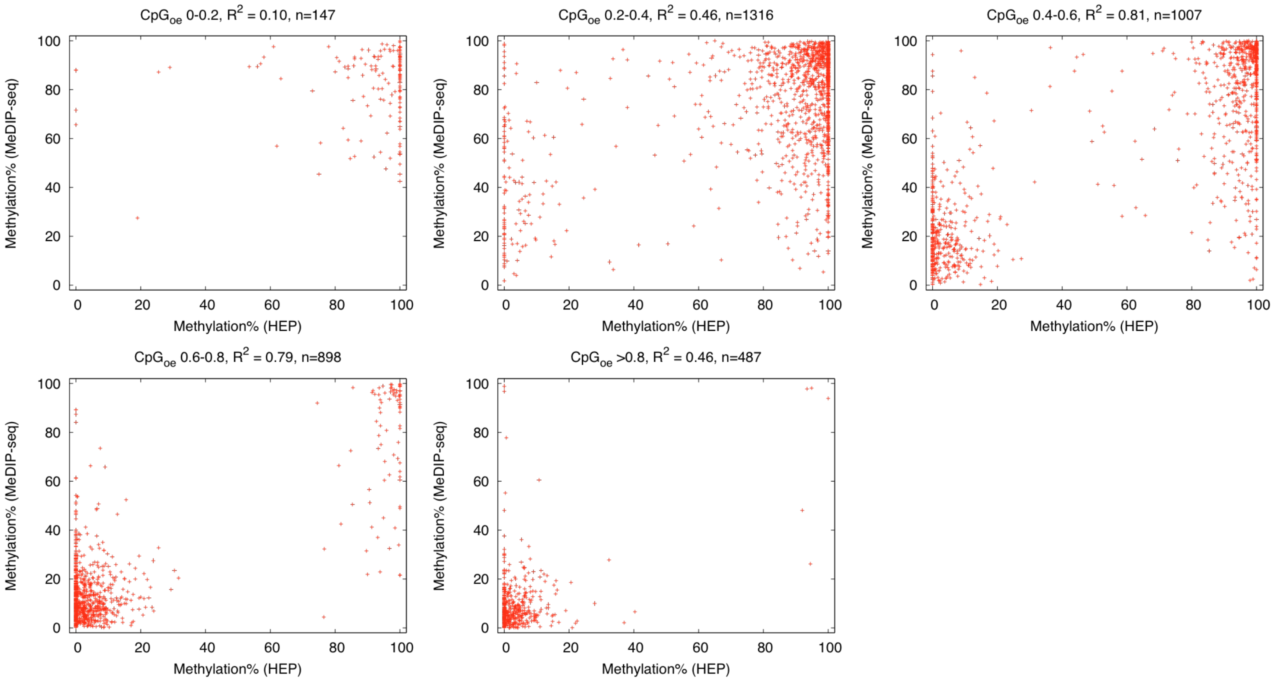
Supplementary Table 3. Regions analyzed in Supplementary Figure 3.

no.	Bisulfite-PCR amplicon ID	Chr	Amplicon start	Amplicon end	GC%	CpG%	Array ROI id	ROI start	ROI end
1	4979	22	29,938,315	29,938,715	67	6.2	27086	29,938,018	29,938,968
2	5105	22	17,545,712	17,546,102	77	11.8	26738	17,545,145	17,547,059
3	5224	22	20,130,777	20,131,135	64	6.7	26831	20,130,463	20,131,012
4	5981	22	38,296,618	38,297,072	67	3.5	27317	38,296,495	38,297,444
5	6135	22	35,970,069	35,970,424	67	4.5	27195	35,970,040	35,970,489
6	6142	22	35,938,425	35,938,903	67	3.8	27194	35,938,422	35,938,871
7	6158	22	49,216,499	49,216,926	61	4.9	27675	49,216,545	49,216,794
8	6313	22	18,510,668	18,511,158	70	8.4	26780	18,510,402	18,511,251
9	6575	22	49,334,595	49,335,038	67	7.4	27696	49,333,374	49,335,023
10	6587	22	29,281,454	29,281,947	63	8.3	27059	29,280,934	29,282,083
11	6696	22	41,419,027	41,419,524	66	8.2	27422	41,418,869	41,419,618
12	6705	22	45,453,093	45,453,592	59	5.6	27559	45,453,377	45,453,526
13	6763	22	39,964,476	39,964,879	70	6.4	27354	39,963,565	39,964,753
14	8828	6	101,018,918	101,019,406	64	6.5	34354	101,018,058	101,020,107
15	9054	6	139,136,417	139,136,888	60	5.3	34696	139,136,090	139,137,339
16	9098	6	46,811,222	46,811,720	51	3.8	34024	46,810,546	46,811,795
17	9106	6	53,322,056	53,322,372	42	2.5	34096	53,320,630	53,322,579
18	9181	6	150,963,434	150,963,699	64	9.8	34785	150,962,740	150,963,841
19	9232	6	28,475,339	28,475,830	59	6.3	33702	28,475,166	28,475,915
20	9253	6	126,111,195	126,111,673	53	4.8	34567	126,110,449	126,113,315
21	9254	6	153,346,203	153,346,693	70	8.0	34814	153,345,105	153,346,654
22	9368	6	170,735,558	170,735,982	51	3.4	35053	170,735,258	170,736,107
23	9480	6	33,787,387	33,787,734	63	8.9	33720	33,787,126	33,787,975
24	9482	6	54,281,191	54,281,533	41	1.2	34106	54,280,991	54,282,009
25	9502	6	154,872,494	154,872,915	67	6.9	34823	154,872,350	154,873,838
26	9520	6	76,368,619	76,368,942	74	13.0	34204	76,367,741	76,369,717
27	9725	6	37,774,253	37,774,700	68	8.3	33827	37,774,107	37,775,056
28	11747	20	2,801,490	2,801,889	57	4.3	24947	2,800,957	2,802,966
29	13405	22	35,777,451	35,777,920	73	10.6	27183	35,777,329	35,778,352

All co-ordinates are based on the NCBI36 version of the human genome
Primer sequences are available upon request

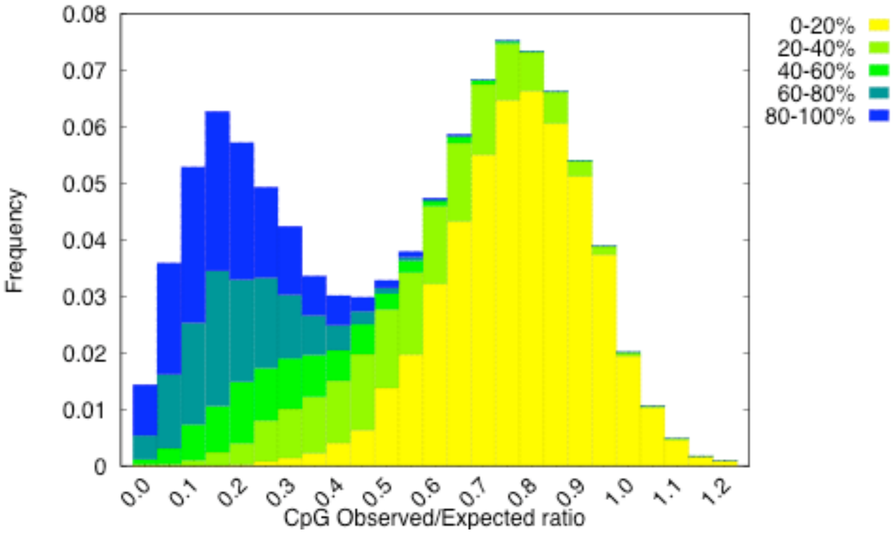


Supplementary Figure 4. Batman analysis of an independently generated MeDIP-chip data set. (A) LOESS-normalized log2 ratios for the MeDIP-chip data corresponding to WI38 fibroblasts[4] was downloaded from GEO, and re-mapped onto the NCBI36 assembly using liftOver data from UCSC. Batman analysis was performed as described in the main text. For each gene with bisulfite sequencing results in Weber et al[4]., we found the TSS in Ensembl, then selected the array probe set within 200bp of that TSS. All but two genes were mapped in this way. We then plotted the bisulfite-PCR sequencing results against either the mean array signal (log_2 ratio) across the probe set, or the mean Batman output across the region. Points are colored according to the promoter classification from Weber et al.[4] (and also indicated in the figure) (B) There were 4 HCP promoters which were reported as being unmethylated by Weber et al., but were called as being substantially methylated by Batman. In all these cases, the Batman output (indicated by gray bars corresponding to 100bp windows across the tiled region) showed distinct methylated and unmethylated regions. The locations of CpG dinucleotidesare are marked with crosses above each plot, with CpGs actually assayed by bisulfite genomic sequencing by Weber et al.[4] highlighted in red. In all four cases, the bisulfite-sequenced region – called as being <20% methylated by Weber et al. – was also called as being <20% methylated by Batman. Hence the apparent discordance observed in Supplementary Figure 4A, is due to the averaging of the Batman results which cover a wider region than that assayed by Weber et al.



Supplementary Figure 5. Comparison of the MeDIP-seq data with bisulfite-PCR sequencing data from the Human Epigenome Project, stratified by CpG density. The data is the same as that in **Figure 4b** of the main text, but split by observed/expected CpG density of 500bp windows centered around the 100 bp tiles (the number of 100 bp tiles that overlap a HEP amplicon in each group is indicated by ‘n’). The distribution of DNA methylation in the human genome is strongly bimodal[3]. Standard correlation measures such as Pearson’s or Spearman’s provide misleading values when there is an insuffіcent range within the dataset. For example, the correlation between MeDIP-Seq and HEP data is obviously very strong in the ‘CpG$_{o/e}$ > 0.8’ category, but this is not reflected in the correlation coeffecient. We therefore quantified the level of agreement between these two datasets by discretizing each measurement into one of three bins: low (0-33%), intermediate (34-66%) and high (67-100%). We counted the fraction of amplicons that fell into the same bin in both the MeDIP-seq and HEP dataset (shown below). This ad-hoc method provides a much better measure of the correlation between MeDIP-Seq and the HEP datasets.

CpGo/e	Agreement between MeDIP-seq and HEP
0.0 - 0.2	76.9%
0.2 - 0.4	70.5%
0.4 - 0.6	76.7%
0.6 - 0.8	91.9%
>0.8	97.3%
Overall	**80.7%**



Supplementary Figure 6. MeDIP-seq DNA methylation profiles of promoters in the human genome in mature sperm. We selected 500bp windows upstream of all TSSs from Ensembl human build 45.36g. For each window that was covered by methylation calls from our MeDIP-seq analysis (as described din the main text), we calculated the mean methylation score and the CpG observed/expected ratio (CpG$_{o/e}$). We plot the distribution of promoter CpG$_{o/e}$ ratios, subdividing each bin according to the range of methylation score.

Supplementary References

1. Weber, M. et al. Nat. Genet. 37, 853-862 (2005).
2. Oberley, M.J. & Farnham, P.J. *Methods Enzymol.* **371**, 577-596 (2003).
3. Slater, G.S. & Birney E. *BMC Bioinformatics* **15**, 31-34 (2005).
4. Weber, M. et al. *Nat. Genet.* **39**, 457-466 (2007).
5. Lewin, J. et al. *Bioinformatics* **20**, 3005-12. (2004)
6. Eckhardt, F. et al. *Nat. Genet.* **38**, 1378-1385 (2006).